

Kronecker-based modeling of networks with unknown communication links ^{*}

B. Sinquin ^{*} M. Verhaegen ^{*}

^{*} Delft Center for Systems and Control
Delft University of Technology
Delft, Netherlands.

Abstract: In this paper we propose a Kronecker-based modeling of large networks with unknown interconnection links. The class of Kronecker networks is defined for which we formulate a Vector Autoregressive model. Its coefficient-matrices are decomposed into a sum of Kronecker products. When the network is labeled such that the number of terms in the sum is small compared to the size of the matrix, exploiting this Kronecker structure leads to high data compression. Two algorithms were designed for an efficient estimation of the coefficient-matrices, namely a non-iterative and overparametrized algorithm as well as an Alternating Least Squares minimization. We prove that the latter always converges to the true parameters for non-zero initial conditions. This framework moreover allows for a convenient integration of more structure (e.g sparse, banded, Toeplitz) on smaller-size matrices. Numerical examples on atmospheric turbulence data has shown comparable performances with the unstructured least-squares estimation while the number of parameters is growing *only linearly* w.r.t. the number of nodes instead of quadratically in the full unstructured matrix case.

Keywords: network modeling, large-scale systems, Kronecker product, low-rank approximation, multi-convex minimization.

1. INTRODUCTION

Modeling large-scale networks has been stirring much developments in various fields such as machine learning and system identification. The assembly of numerous systems interacting with one another arises in fields such as biology, e.g with the brain neurons in Bullmore and Sporns (2009), optics with the atmospheric turbulence, and many others. Due to the large size of input-output data batches, identifying locally the behavior of the network is a major challenge that has been mainly addressed by using prior knowledge on how the subsystems, or nodes, are connected to one another. One common assumption is sparsity and relies on the fact that each node is connected to a limited number of other nodes with respect to the network's size. Other well-studied structures include interconnected one-dimensional strings of subsystems in Rice (2010), or clusters of different subsystems with known connection patterns, named as alpha-heterogeneous in Massioni (2014). However the links between the subsystems in the network might not be known beforehand. In the so-called sparse plus low rank networks -Zorzi and Chiuso (2016)-, a few latent variables relate most of the measured nodes from which few of them influence each other. Model identification remains computationally challenging to handle the combination of these two matrices structure. In Leskovec et al. (2010) very general graphs are studied, whose weighted adjacency matrix is approximated with Kronecker products of a so-called initiator matrix. Such an operation replicates the network structure associated to the initiator matrix

to higher dimensions. Therefore this matrix embeds all the required information on the network to construct it at higher scales. It is moreover shown in Leskovec et al. (2010) that any network is approximated with such a structure after having ordered the nodes in the most adequate way.

In this paper we address the identification of (2D) spatial-temporal dynamical models of the Vector-Auto-Regressive (VAR) form. The coefficient-matrices of this model are parametrized as sums of Kronecker products. Loan and Pitsianis (1992) establishes the equivalence between expressing a matrix as a sum containing few Kronecker products and a low-rank approximation of a reshuffled matrix. The latter has been studied in Tyrtyshnikov (2004) for function-related matrices, in which more general existence theorems are derived. Moreover, when the matrix exhibits multiple symmetries or is block-Toeplitz with Toeplitz-blocks, it is guaranteed that such a low rank approximation of the reshuffled matrix exists, see Loan and Vokt (2015) and Kamm and Nagy (2000). More than only enjoying the storage of a reduced number of entries, such a structure enables fast computations thanks to the very pleasant algebra of the Kronecker product, Loan (2000). For a matrix written as $A = A_1 \otimes A_2$ with $A_1, A_2 \in \mathbb{R}^{N \times N}$, matrix-matrix multiplication and inversion both require $\mathcal{O}(2N^3 + N^4)$ instead of $\mathcal{O}(N^6)$ for the unstructured case.

For a given labeling of the network, the coefficient-matrices in the VARX model may be represented with as few terms as possible in the Kronecker sum while still guaranteeing a given level of performance with respect to the standard least squares solution with unstructured coefficient matrices. A major challenge in the estimation is the computational efficiency. We address this problem

^{*} Corresponding author: b.sinquin@tudelft.nl.

This work is sponsored by the European Research Council, Advanced Grant Agreement No. 339681.

Work presented at the ERNSI 2016 Workshop, Italy

by parametrizing only the factor matrices, e.g. A_1, A_2 , which gives rise to a bilinear least squares problem. The estimation problem belongs to the class of multi-convex optimization: fixing all variables but one yields a convex problem, Shen et al. (2016). Such a formulation shares similarities with the identification of Hammerstein systems, see e.g. Wang (2009) for which a two-stage algorithm is proposed. The contribution of this work includes the formulation of a Kronecker-based VARX model for networks with unknown communication links. The estimation problem is solved using both a non-iterative three-stage algorithm and an iterative Alternating Least Squares, for which we have extended a convergence proof to our case. Thirdly, the missing sensor data can be retrieved by formulating a multi-linear least squares.

This paper is organized as follows. In the second section the class of *Kronecker networks* is defined. The third section formulates the Kronecker VARX identification framework while the fourth section describes the non-iterative overparametrized algorithm to estimate the factor matrices with minimum computational complexity. An alternative is proposed in Section V within the framework of multi-linear least squares. Section VI describes the missing data case. This covers the case of non-rectangular measurement grids. We study how additional structure can be considered on the factor matrices in Section VII. Last, Section VIII is dedicated to numerical experiments.

Notations. The vectorization operator for a matrix X is $\text{vec}(X)$. $\text{ivec}(x)$ reshapes the vector x into a matrix whose size will be clear from the context. The Kronecker product between two matrices X, Y is denoted by $X \otimes Y$. The 2-norm of a vector x is written as $\|x\|_2$. The number of non-zero-elements in a vector x is $\|x\|_0$ while the sum in absolute value is $\|x\|_1$. $\lambda_{\max}(X)$ is the largest eigenvalue of the positive-semidefinite matrix X . The nuclear norm of X , denoted with $\|X\|_*$, represents the sum of the singular values of X .

2. PRELIMINARIES

Definition 1. (Loan (2000)). Let X be a $m_1 \times n_1$ block matrix with blocks $X(i, j)$ in $\mathbb{R}^{m_2 \times n_2}$, given as:

$$X = \begin{bmatrix} X(1,1) & \cdots & X(1,2) \\ \vdots & \ddots & \vdots \\ X(m_1,1) & \cdots & X(m_1,n_1) \end{bmatrix} \in \mathbb{R}^{m_1 m_2 \times n_1 n_2}$$

then the re-shuffle operator $\mathcal{R}(X)$ is defined as:

$$\mathcal{R}(X) = \begin{bmatrix} \text{vec}(X(1,1))^T \\ \vdots \\ \text{vec}(X(m_1,1))^T \\ \text{vec}(X(1,2))^T \\ \vdots \\ \text{vec}(X(m_1,2))^T \\ \vdots \\ \text{vec}(X(m_1,n_1))^T \end{bmatrix} \in \mathbb{R}^{m_1 n_1 \times m_2 n_2}$$

Reshuffling $Y = \mathcal{R}(X)$ to form X back is defined with the operator \mathcal{P} , i.e. $\mathcal{P}(Y) = X$.

Lemma 1. (Loan (2000)). Let X be defined as in Definition 1, and let $X = F \otimes G$, with $F, G \in \mathbb{R}^{m_1 \times n_1} \times \mathbb{R}^{m_2 \times n_2}$. Then:

$$\mathcal{R}(X) = \text{vec}(F)\text{vec}(G)^T$$

The operation in Lemma 1 can also be reversed by the definition of the inverse vec operator $\text{ivec}(\cdot)$.

Lemma 2. (Loan (2000)). Let X be defined as in Definition 1, and let an SVD of $\mathcal{R}(X)$ be given as:

$$\mathcal{R}(X) = \sum_{\ell=1}^r \sigma_\ell u_\ell v_\ell^T$$

and let $\text{ivec}(u_\ell) = U_\ell$, $\text{ivec}(v_\ell) = V_\ell$:

$$X = \sum_{\ell=1}^r \sigma_\ell U_\ell \otimes V_\ell$$

The integer r is called the *Kronecker rank* of X w.r.t. the chosen block partitioning of X as given in definition 1.

Definition 2. (α -decomposable matrices, Massioni (2014)).

Let P be a $N \times N$ pattern matrix. Define $\beta_j = \sum_{i=1}^j N_i$ (with $\beta_0 = 0$) and $I_{[a_1:a_2]}$ as an $N \times N$ diagonal matrix which contains 1 in the diagonal entries of indices from a_1 to a_2 (included) and 0 elsewhere, then an α -decomposable matrix (for a given α) is a matrix of the following kind:

$$\mathcal{M} = \sum_{i=1}^{\alpha} (I_{[\beta_{i-1}+1:\beta_i]} \otimes M_a^{(i)}) + \sum_{i=1}^{\alpha} (I_{[\beta_{i-1}+1:\beta_i]} P \otimes M_b^{(i)})$$

The matrices $M_a^{(i)}$ are the diagonal blocks of \mathcal{M} , while the matrices $M_b^{(i)}$ constitute the off-diagonal blocks, according to the structure of \mathcal{P} .

For $\alpha = 1$ (and $\theta_1 = N$), these matrices are simply called *decomposable* matrices. The class of α -decomposable matrices will be denoted by \mathcal{D}^α , with for $\alpha = 1$ just the symbol \mathcal{D} will be used.

As a generalization of this class of structured matrices, we define next the class of sums of Kronecker product matrices.

Definition 3. The class of sums of Kronecker product matrices, denoted by \mathcal{S} , contains matrices of the following kind:

$$\mathcal{M} = \sum_{i=1}^r M_a^{(i)} \otimes M_b^{(i)}$$

with $M_a^{(i)} \in \mathbb{R}^{m_1 \times n_1}$ and $M_b^{(i)} \in \mathbb{R}^{m_2 \times n_2}$

3. PROBLEM FORMULATION

The sensor readings at time instance k are stored in the matrix $S(k)$ as:

$$S(k) = \begin{bmatrix} s_{1,1}(k) & s_{1,2}(k) & \cdots & s_{1,N}(k) \\ s_{2,1}(k) & s_{2,2}(k) & \cdots & s_{2,N}(k) \\ \vdots & \vdots & \ddots & \vdots \\ s_{N,1}(k) & s_{N,2}(k) & \cdots & s_{N,N}(k) \end{bmatrix} \quad (1)$$

with $s_{i,j}(k) \in \mathbb{R}$. In this paper we will consider that the (temporal) dynamics of this array of sensors is governed by the following VAR(X) model:

$$\text{vec}(S(k)) = \sum_{i=1}^p A_i \text{vec}(S(k-i)) + C_0 \text{vec}(E(k)) \quad (2)$$

with $\text{vec}(E(k))$ zero-mean white noise with covariance matrix I . The coefficient matrices A_i and C_0 in the VARX (we will restrict for simplicity to the AR-case) model (2), in general are highly structured. Here we will consider these coefficient matrices to be in the matrix sets \mathcal{D}^α , \mathcal{D} or \mathcal{S} . We will consider the case they belong to the set \mathcal{S} , and for the moment only focus on the coefficient matrices A_i . To address an identification problem we will parametrize these coefficient matrices as:

$$A_i = \sum_{j=1}^{r_i} M(b_i^{(j)})^T \otimes M(a_i^{(j)}) \quad (3)$$

with the vectors $a_i^{(j)}$ and $b_i^{(j)}$ parametrizing the matrices $M(a_i^{(j)})$ and $M(b_i^{(j)})$ in an affine manner. If we consider the term $C_0 \text{vec}(E(k))$ as a temporally white sequence $v(k)$, then the ARX model (2) can be written as,

$$\text{vec}(S(k)) = \sum_{i=1}^p \left(\sum_{j=1}^{r_i} M(b_i^{(j)})^T \otimes M(a_i^{(j)}) \right) \text{vec}(S(k-i)) + v(k) \quad (4)$$

Using the following Kronecker rule, for matrices X, Y, Z of compatible dimensions such that the product XYZ exists,

$$\text{vec}(XYZ) = (Z^T \otimes X) \text{vec}(Y) \quad (5)$$

we can write the ARX model (4) as,

$$S(k) = \sum_{i=1}^p \left(\sum_{j=1}^{r_i} M(a_i^{(j)}) S(k-i) M(b_i^{(j)}) \right) + V(k) \quad (6)$$

with $\text{vec}(V(k)) = v(k)$. This can also be written explicitly as,

$$S(k) = \sum_{i=1}^p \left[M(a_i^{(1)}) \dots M(a_i^{(p)}) \right] (I_{r_i} \otimes S(k-i)) \begin{bmatrix} M(b_i^{(1)}) \\ \vdots \\ M(b_i^{(p)}) \end{bmatrix} + V(k)$$

The AR(X) models (4), (6) or (7) are called *Kronecker ARX network models*, or briefly *Kronecker ARX models* (KrARX) (pronounced as "quarks").

3.1 The identification problem of KrARX models.

Given the model structure of KrARX models, the problem of identifying these models from measurement sequences $\{S(k)\}_{k=1}^{N_t}$ is fourfold:

- (1) The temporal order index p .
- (2) The spatial order index r_i for each coefficient matrix.
- (3) The parametrization of the matrices $M(a_i^{(j)})$ and $M(b_i^{(j)})$. An example of a parametrization of the matrices $M(a_i^{(j)})$ and $M(b_i^{(j)})$ is (block) Toeplitz.
- (4) The estimation of the parameter vectors $a_i^{(j)}$, $b_i^{(j)}$ up to an ambiguity transformation. This requires the specification of a cost function. An example of such a cost function using the model (6) is the following least squares cost function,

$$\min_{p, r_i, a_i^{(j)}, b_i^{(j)}} \sum_{k=p+1}^{N_t} \left\| S(k) - \sum_{i=1}^p \left(\sum_{j=1}^{r_i} M(a_i^{(j)}) S(k-i) M(b_i^{(j)}) \right) \right\|_F^2 \quad (8)$$

for data batches with N_t temporal samples. By the selection of the parameter p and particular choices of the parametrizations in step 3 above, various special cases of restricting the coefficient matrices A_i in (2) to particular sets (such as \mathcal{D}^α , \mathcal{D} or \mathcal{S}) can be considered. Further constraints to the (least-squares) cost function, such as in (8), might be introduced to look for sparsity in the parametrization vectors $a_i^{(j)}$ and $b_i^{(j)}$.

An important challenge of the parameter estimation problem is the *computational efficiency*. The covariance matrix estimation in high dimensional spaces has already been addressed in Tsiligkaridis and Hero (2013) and is not considered further on in this document.

4. ESTIMATING KRARX MODELS WITH A THREE-STAGE APPROACH

4.1 A global least squares cost function with rank minimization

Consider the KrARX model (4) then we can define the matrix Θ_i with $\left(\sum_{j=1}^{r_i} M(b_i^{(j)})^T \otimes M(a_i^{(j)}) \right)$ and write this model as:

$$\text{vec}(S(k)) = \sum_{i=1}^p \Theta_i \text{vec}(S(k-i)) + v(k) \quad (9)$$

According to (4) and the definition of the re-shuffling operator $\mathcal{R}(\cdot)$ we have,

$$\mathcal{R}(\Theta_i) = \sum_{j=1}^{r_i} \text{vec}(M(b_i^{(j)})^T) \text{vec}(M(a_i^{(j)}))^T$$

Therefore a way to find the spatial order index (assuming it is the same for all Θ_i) is via the Kronecker rank. Let this be denoted by r , then we write,

$$\mathcal{R}(\Theta_i) = \sum_{j=1}^r \sigma_j^i u_j^i (v_j^i)^T$$

and with the definition of $\text{ivec}(u_j^i) = U_j^i$, $\text{ivec}(v_j^i) = V_j^i$, the coefficient matrix Θ_i can be written as,

$$\Theta_i = \sum_{j=1}^r \sigma_j^i U_j^i \otimes V_j^i \quad (10)$$

Not knowing the Kronecker rank, a possible way to retrieve this parameter and the coefficient matrices Θ_i for a given temporal order p from the data is via the following multi-criteria cost function:

$$\min_{\Theta_i} \sum_{k=p+1}^{N_t} \left\| \text{vec}(S(k)) - \sum_{i=1}^p \Theta_i \text{vec}(S(k-i)) \right\|_F^2 \quad (11)$$

$$\min_{\Theta_i} \text{rank}(\mathcal{R}(\Theta_i))$$

Let the estimated coefficient matrices be denoted by $\hat{\Theta}_i$, then subsequently an SVD of the matrices $\mathcal{R}(\hat{\Theta}_i)$ provides estimates of the terms σ_j^i , U_j^i and V_j^i in the Kronecker products in (10).

It should be remarked that this way of formulating the identification problem does not require a parametrization of the matrices $M(a_i^{(j)})$ and $M(b_i^{(j)})$ as stipulated in the

third step of the fourfold generic identification problem formulation outlined in Section 3.1.

Since the rank operator in the cost function (11) turns this cost function into a non-convex optimization problem, the nuclear norm can be used to convexify this problem. This would then yield the following problem formulation,

$$\min_{\Theta_i} \sum_{k=p+1}^{N_t} \|\text{vec}(S(k)) - \sum_{i=1}^p \Theta_i \text{vec}(S(k-i))\|_F^2 + \lambda \sum_{i=1}^p \|\mathcal{R}(\Theta_i)\|_* \quad (12)$$

where an additional weighting parameter λ is introduced. The nuclear norm regularization on p matrices of size $N^2 \times N^2$ is prohibitive especially when handling large datasets. It would indeed imply solving e.g a Alternating Direction Method of Multipliers Algorithm (ADMM) with a singular-value decomposition on large matrices at each iteration or expensive matrices inversions scaling up to $\mathcal{O}(N^6)$.

However we see in the following section that some more efficient computations can be performed by parallelizing the optimization problem.

4.2 Local least squares for parallel computations

In this paragraph we consider $p = 1$ for the sake of clarity. The least squares term in the cost function (12) can be addressed row by row. The ℓ -th line of Θ_1 , where $\ell = N(n-1) + j$, and j, n integers, is denoted with $\theta_{n,j}$. Using "standard" matlab notation to select part of a matrix, the matrix $\mathcal{P}(\theta_{i,j})$ and $\mathcal{R}(\Theta_1)$ are related as:

$$\mathcal{P}(\theta_{i,j})^T = \mathcal{R}(\Theta_1)(i : N : \text{end}, j : N : \text{end}) \quad (13)$$

To further clarify this notation, we refer to Figure 1 for a display of the different matrices in the above relation. We

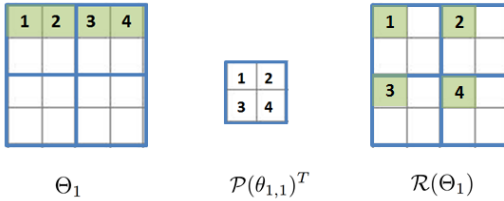


Fig. 1. Schematic representation of the matrices $\Theta_1, \mathcal{P}(\theta_{1,1})^T$ and $\mathcal{R}(\Theta_1)$.

state the following lemma:

Lemma 3. If $\text{rank}(\mathcal{R}(\Theta_1)) = r < N$, then for all $i, j = 1..N$, $\text{rank}(\mathcal{P}(\theta_{i,j})^T) \leq r$.

Proof. Let the PCA decomposition of $\mathcal{R}(\Theta_1)$ be such that:

$$\mathcal{R}(\Theta_1) = \begin{bmatrix} u_1(1,:) \\ \vdots \\ u_N(1,:) \\ \vdots \\ u_1(N,:) \\ \vdots \\ u_N(N,:) \end{bmatrix} \begin{bmatrix} v_1(1,:)^T & \dots & v_N(1,:)^T & \dots & v_N(N,:)^T \end{bmatrix}^T \quad (14)$$

where $u_i, v_j \in \mathbb{R}^{N \times r}$. From (13), $\mathcal{P}(\theta_{i,j})^T = u_i v_j^T$. The vectors u_i, v_j are not necessarily full column rank, therefore $\text{rank}(\mathcal{P}(\theta_{i,j})^T) \leq r$.

The reverse implication is not true. Denote an upper-bound on the rank of $\mathcal{R}(\Theta_1)$ by r_{max} . By assuming that the PCA decomposition of $\mathcal{P}(\theta_{i,j})^T$ is given with $u_i v_j^T$ and $u_i, v_j \in \mathbb{R}^{N \times r_{max}}$, a low-rank matrix $\mathcal{R}(\Theta_1)$ is built using (13).

We describe the algorithm in the following lines. Let the measurement at position (j, n) in the data matrix $S(k)$ correspond to the ℓ -th entry of the vector $\text{vec}(S(k))$. Then a rank-constrained least squares optimization is formulated to estimate the matrix $\Theta_1(\ell, :)$ without over-fitting:

$$\min_{\Theta_1(\ell,:)} \sum_{k=p+1}^{N_t} \|\text{vec}(S(k))(\ell, :) - \Theta_1(\ell, :)\text{vec}(S(k-1))\|_2^2 \quad \text{s.t. } \text{rank}(\mathcal{P}(\Theta_1(\ell, :))) = r_{max}$$

which is solved using the relaxed problem:

$$\min_{\Theta_1(\ell,:)} \sum_{k=p+1}^{N_t} \|\text{vec}(S(k))(\ell, :) - \Theta_1(\ell, :)\text{vec}(S(k-1))\|_2^2 + \lambda \|\mathcal{P}(\Theta_1(\ell, :))\|_* \quad (14)$$

A PCA of the low-rank matrix $\mathcal{P}(\Theta_1(\ell, :))^T$ yields the following decomposition:

$$\mathcal{P}(\Theta_1(\ell, :))^T = \hat{u}_n \hat{v}_j^T \quad (15)$$

where $\hat{u}_n, \hat{v}_j \in \mathbb{R}^{N \times r}$. This decomposition is unique up to a (non-singular) ambiguity transformation $T \in \mathbb{R}^{r \times r}$:

$$\mathcal{P}(\Theta_1(\ell, :))^T = u_n T T^{-1} v_j = \hat{u}_n \hat{v}_j^T$$

Therefore such a PCA cannot be performed for N independent well-chosen rows, as it would yield N different ambiguity transformations. There remains $2(N-1)$ matrices of size $N \times r$ to estimate the full factor matrices \hat{u}, \hat{v} . The latter have to be consistent with the estimation in (15) and consider the same ambiguity transformation. Therefore, if (14) is solved e.g for $\ell = 1$, then for all $\ell \in [2, N]$, we solve the constrained least-squares optimization:

$$\min_{\Theta_1(\ell,:), \hat{u}_n} \sum_{k=p+1}^{N_t} \|\text{vec}(S(k))(\ell, :) - \Theta_1(\ell, :)\text{vec}(S(k-1))\|_2^2 \quad \text{s.t. } \mathcal{P}(\theta_{n,1})^T = \hat{u}_n \hat{v}_1^T \quad (16)$$

and for all ℓ such that $\ell = N(j-1) + 1$, where $j \in [2, N]$:

$$\min_{\Theta_1(\ell,:), \hat{v}_\ell} \sum_{k=p+1}^{N_t} \|\text{vec}(S(k))(\ell, :) - \Theta_1(\ell, :)\text{vec}(S(k-1))\|_2^2 \quad \text{s.t. } \mathcal{P}(\theta_{1,j})^T = \hat{u}_1 \hat{v}_j^T \quad (17)$$

These $2(N-1)$ least squares can be performed in parallel, each of which corresponds to one sensor location as can

be visualized in Figure 2. Choosing the sensor location in position (1, 1) in (14) is not unique.

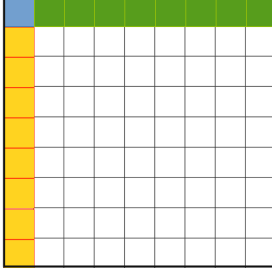


Fig. 2. Rectangular array: the colored entries locate in the matrix $S(k)$ which data to consider in order to estimate the coefficient-matrices with minimum computational complexity. Blue entry: minimization of type (23)-(15). Green entry: minimization of type (16). Yellow entry: minimization of type (17).

The three-step algorithm has been discussed for the case $r < N$. In most cases, the Kronecker rank is not known *a priori* and it has to be detected with cross-validation. Therefore we analyze how to deal with higher Kronecker ranks, e.g. $r < 2N$. In the previous paragraph, the Kronecker rank was limited by the size of the submatrices $\mathcal{P}(\theta_{i,j}) \in \mathbb{R}^{N \times N}$. Therefore a submatrix of size $2N \times 2N$ shall be selected such that the PCA in (15) is then carried out on a rank-deficient matrix. For example, the output data associated with the set of indices $\mathcal{L} = \{1, 2, N+1, N+2\}$ conveys enough information to retrieve the factor matrices in this case. Figure 3 illustrates the sensor locations. The regularized least squares in (14) is then extended into:

$$\min_{\Theta_1(\ell,:)} \sum_{k=p+1}^{N_t} \sum_{\ell \in \mathcal{L}} \|\text{vec}(S(k))(\ell, :) - \Theta_1(\ell, :) \text{vec}(S(k-1))\|_2^2 + \lambda \|\mathcal{R}_{\mathcal{L}}\|_* \quad (18)$$

where the rank-deficient matrix is:

$$\mathcal{R}_{\mathcal{L}} = \begin{bmatrix} \mathcal{P}(\theta_{1,1})^T & \mathcal{P}(\theta_{1,2})^T \\ \mathcal{P}(\theta_{2,1})^T & \mathcal{P}(\theta_{2,2})^T \end{bmatrix}$$

Similarly, the PCA is performed on $\mathcal{R}_{\mathcal{L}}$ and the least squares in (16) and (17) are formulated for 4 neighboring points.

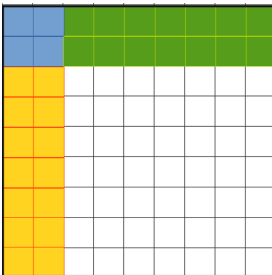


Fig. 3. Identification for matrices of Kronecker rank between N and $2N$. The color code is the same as in Figure 2.

The computational complexity is reduced exploiting the Kronecker structure, and is attractive for being non-iterative and parallelizable to a large extent. Although it

is possible to enforce some additional structure, e.g. block-Banded with banded-Blocks, the framework is nonetheless less elegant than the Alternating Least Squares minimization that we propose in the next paragraph to embed more structure on the factor matrices.

5. BI-CONVEX COST FUNCTION APPROACH

5.1 Alternating Least Squares

This approach is formulated based on the representation (7). This forms has the advantage that constraints on the parametrizations of the matrices $M(a_i^{(j)})$ and $M(b_i^{(j)})$ can be more easily taken into consideration. For example the parametrization may enforce the matrices to have a (block) Toeplitz structure with entries in a particularly chosen set, e.g. $\{0, 1\}$. Denote this set for the parameter vectors $a_i^{(j)}$, $b_i^{(j)}$ resp. by \mathcal{T}_a and \mathcal{T}_b and let us assume p and r to be given. Denote moreover:

$$\tilde{M}_{a_i} := \begin{bmatrix} M(a_i^{(1)}) & \cdots & M(a_i^{(r)}) \end{bmatrix}$$

$$\tilde{M}_{b_i} := \begin{bmatrix} M(b_i^{(1)}) \\ \vdots \\ M(b_i^{(p)}) \end{bmatrix}$$

Then we have the following bi-convex optimization problem

$$\min_{a_i^{(j)} \in \mathcal{T}_a, b_i^{(j)} \in \mathcal{T}_b} \sum_{k=p+1}^{N_t} \|S(k) - \sum_{i=1}^p \tilde{M}_{a_i} (I_r \otimes S(k-i)) \tilde{M}_{b_i}\|_F^2 \quad (19)$$

Here again we consider $p = 1, r = 1$ for the sake of clarity. The KrARX model then reads:

$$S(k) = LS(k-1)R + V(k) \quad (20)$$

such that $A_1 = R^T \otimes L$. The i -th column of L , resp. R , is denoted with ℓ_i , resp. r_i . Given an initial guess of the matrix L denoted as $\hat{L} := L_0$, the following three steps are performed:

- (1) *Step 1:* A minimization step over the columns of the matrix R :

$$\forall j = 1..N, \quad \min_{r_j} \sum_{k=1}^{N_t} \|S(k)(:, j) - \hat{L}S(k-1)r_j\|_2^2 \quad (21)$$

- (2) *Step 2:* the estimates r_j are normalized:

$$\hat{r}_j = \frac{r_j}{\|r_j\|_2} \quad (22)$$

- (3) *Step 3:* A minimization step over the columns of the matrix L :

$$\forall j = 1..N, \quad \min_{\ell_j} \sum_{k=1}^{N_t} \|S(k)(j, :) - \ell_j^T S(k-1)\hat{R}\|_2^2 \quad (23)$$

5.2 Convergence proof for the Alternating Least Squares

In this section we study the convergence of the alternating least squares repeating the three steps in (21)-(23). The proof relies on the work in Li et al. (2004) that establishes the result when L, R are vectors. Therefore we review here the main theorems of the proof, and highlight the changes

in Appendix. We first reformulate the equation (20). It can be shown that:

$$\begin{aligned}\tilde{\mathbf{S}} &= M_r L + \tilde{\mathbf{V}} \\ \tilde{\mathbf{S}} &= M^\ell R + \tilde{\mathbf{V}}\end{aligned}$$

where

$$\begin{aligned}\tilde{\mathbf{S}} &= \begin{bmatrix} \tilde{S}(1,1) & \dots & \tilde{S}(N,1) \\ \vdots & & \vdots \\ \tilde{S}(1,N) & \dots & \tilde{S}(N,N) \end{bmatrix} \in \mathbb{R}^{N_t N \times N} \\ \tilde{S}(j,i) &= \begin{bmatrix} S_1(j,i) \\ \vdots \\ S_{N_t}(j,i) \end{bmatrix} \\ M_r &= (I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} I_N \otimes r_1 \\ \vdots \\ I_N \otimes r_N \end{bmatrix} \\ M^\ell &= (I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} \ell_1 \otimes I_N \\ \vdots \\ \ell_N \otimes I_N \end{bmatrix} \\ \tilde{\mathbf{U}} &= \begin{bmatrix} U_1(1,:) & \dots & U_1(N,:) \\ \vdots & & \vdots \\ U_{N_t}(1,:) & \dots & U_{N_t}(N,:) \end{bmatrix} \in \mathbb{R}^{N_t \times N^2}\end{aligned}$$

$\tilde{\mathbf{V}}$ is defined similarly as $\tilde{\mathbf{S}}$. Li et al. (2004) analyse the convergence of the Alternating Least squares solution using the Contraction Mapping Theorem, Granas and Dugundji (2001). Let κ denote the iteration counter. Let the initial estimate of \hat{L} be denoted by $\hat{L}(\kappa)$ and denote the results of the subsequents three steps (1 to 3) be denoted resp. $\hat{R}(\kappa), \hat{R}_{op}(\kappa), \hat{L}(\kappa+1)$ then a *functional* representation of the three steps reads:

$$\begin{aligned}\hat{R}_{op}(\kappa) &= \mathcal{F}_1(\hat{L}(\kappa)) = \left((M^{\hat{L}(\kappa)})^T M^{\hat{L}(\kappa)} \right)^{-1} (M^{\hat{L}(\kappa)})^T \tilde{\mathbf{S}} \\ \hat{R}(\kappa) &= \mathcal{F}_2(\hat{R}_{op}(\kappa)) \\ \hat{L}(\kappa+1) &= \mathcal{F}_3(\hat{R}(\kappa)) = \left(M_{\hat{R}(\kappa)}^T M_{\hat{R}(\kappa)} \right)^{-1} M_{\hat{R}(\kappa)}^T \tilde{\mathbf{S}}\end{aligned}$$

These equations can be expressed using a single operator $\mathcal{F}(\cdot)$ mapping the estimate $\hat{L}(\kappa)$ to $\hat{L}(\kappa+1)$:

$$\begin{aligned}\hat{L}(\kappa+1) &= \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(\hat{L}(\kappa)))) \\ &= \mathcal{F}(\hat{L}(\kappa))\end{aligned}$$

Theorem 4. [The Contraction Mapping Theorem, Granas and Dugundji (2001)] Let (X, D) be a non-empty complete metric space where D is a metric on X . Let $\mathcal{F} : X \rightarrow X$ be a contraction mapping on X , i.e., there is a nonnegative real number $Q < 1$ such that $D(\mathcal{F}(x), \mathcal{F}(y)) \leq QD(x, y)$, for all $x, y \in X$. Then the map \mathcal{F} admits one and only one fixed point $x^* \in X$ which means $x^* - \mathcal{F}(x^*) = 0$. Furthermore, this fixed point can be found from the convergence of an iterative sequence defined by $x(\kappa+1) = \mathcal{F}(x(\kappa))$ for $k = 1, 2, \dots$ with an arbitrary starting point $x(0)$ in X .

We start by defining the following inner product.

Definition 4. Let $X, Y \in \mathbb{R}^{N \times N}$ and denote their columns with x_i, y_i . For two matrices \mathbf{X}, \mathbf{Y} of conformable sizes,

such that $\mathbf{X} = (I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} I \otimes x_1 \\ \vdots \\ I \otimes x_N \end{bmatrix}$, similarly for \mathbf{Y} , the inner product on $\mathbb{R}^{N N_t \times N}$ is defined with:

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \lambda_{max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}}) \|\text{vec}(X)\|^T \text{vec}(Y)\|_2$$

The matrices M_r, M^ℓ have the structure of the matrices \mathbf{X}, \mathbf{Y} in the above definition.

Lemma 5. For the matrix \mathbf{X} and the defined inner product in Definition 4, the quantity $\|\mathbf{X}\|_2 = \sqrt{\langle \mathbf{X}, \mathbf{X} \rangle}$ is a norm on $\mathbb{R}^{N N_t \times N}$.

We define now two sets, associated to each of the variables R and L :

$$\begin{aligned}X_R &= \{ \hat{R} \in \mathbb{R}^{N \times N} \mid \|\hat{r}_i\|_2 = \|r_i\|_2, \quad \hat{r}_1(1) > 0 \} \\ X_L &= \{ \hat{L} \in \mathbb{R}^{N \times N} \mid \|\hat{\ell}_i\|_2 \leq \|\ell_i\|_2 \}\end{aligned}$$

$\hat{R}, \hat{R}_{op}, \hat{L}$ are the estimates of resp. R, R_{op}, L .

Theorem 6. If the following statements are true:

- **A1:** the noise components in V are independent identically distributed (i.i.d) with zero-mean and finite temporal and spatial variance.
- **A2:** the matrix $\tilde{\mathbf{U}}$ is full column rank, which corresponds to the temporal persistency of excitation.
- **A3:** either $\|r_i\|_2$ or $\|\ell_i\|_2$ is known for all i , and the first non-zero entry of r_1 is strictly positive.
- **A4:** the initial estimate L_0 is non-zero.

Then, the map $\mathcal{F} : X_L \rightarrow X_L$ is a contraction on X_L and has a unique fixed point $\hat{L} \in X_L$ when $N_t \rightarrow \infty$, that corresponds to the true parameters.

This theorem proves that, whatever the (non-zero) initial conditions are, the alternating least squares in equations (23)-(21) converges to a global minimum when $N_t \rightarrow \infty$. That global minimum corresponds to the true parameters of the KrARX model.

5.3 Computational complexity

The analysis focuses on one alternating step because the cost is identical for the other minimization step. Computing M_r costs $\mathcal{O}(N_t N^3)$. At each iteration, the least-squares estimate of L is computed as $(M_r^T M_r)^{-1} M_r^T \tilde{\mathbf{S}}$. The matrix-matrix multiplication is done in parallel by sending to N^2 cores the vector $M_r(:, i)$, each of which processes $M_r(:, i)^T M_r(:, i)$ with $\mathcal{O}(N_t)$. A Cholesky factorization is then computed with $\mathcal{O}(N^3/3)$, hence enabling faster inversion with Gaussian pivoting on triangular matrices. Denote $\alpha = (M_r^T M_r)^{-1}$. The global cost for computing the matrix-matrix multiplication $\alpha M_r^T (:= \beta)$ is $\mathcal{O}(N^3 N_t)$, which can be reduced by distributing the computations over different cores. Last, $\beta \tilde{\mathbf{S}}$ requires $\mathcal{O}(N^2 N_t)$. A total of $\mathcal{O}(2N^3 N_t)$ operations are needed to solve one alternating minimization. Hence the ALS requires $\mathcal{O}(4N_{iter} N^3 N_t)$ FLOPS, where N_{iter} is the total number of iterations. Solving the unstructured least squares requires $\mathcal{O}(N^6)$. More importantly, part of the bottleneck resides into matrix-matrix multiplications for which Graphical Processing Units (GPU) are efficient computing tools.

6. MISSING DATA

We extend here the framework considered in (1) that collects the sensor data in a matrix. Let us assume that there are a few missing entries in the map $S(k)$, $k = 1..N_t$. For example, the data might be collected on a circular array, which implies that the blocks in the matrices don't share the same size. Consequently, the coefficient-matrices in the lifted VARX model do not retain the Kronecker structure. Therefore we embed the network in a rectangular envelope, with the added entries considered as unknowns. In this paragraph we study the estimation of the factor matrices as well as the missing entries by working on the rectangular embedding.

Denote the set of known, resp. unknown, entries in $S(k)$, $k = 1..N_t$ with Ω , resp. \mathcal{U} . The global least squares boils down to:

$$\sum_{k=2}^{N_t} \sum_{(i,j) \in \Omega} \|s_{i,j}(k) - \ell_i^T S(k-1) r_j\|_2^2 \quad (24)$$

The estimation problem still belongs to the class of multi-convex problems, Shen et al. (2016). The alternating minimization algorithm consists of minimizing each set of variables while the others are fixed and iterate until convergence. For example, we start by initializing R_0 and $s_{i,j}(k-1)$, for all $i, j \in \mathcal{U}$, and optimize over ℓ_i . Eventually this will provide with estimates of the unknown sensor measurements for all $k < N_t$.

7. STRUCTURED FACTOR MATRICES

The parametrization of the factor matrices based on additional knowledge of the network may help either to reduce the computational complexity of the model identification step, or to cast the model into a structure useful for future use, e.g control. The first category include banded, symmetric, Toeplitz and circulant patterns whereas the second contains e.g sparse or Sequentially Semi-Separable (SSS) matrices. Exploring such structures on the factor matrices is very attractive numerically as the problem size reduces further.

7.1 Toeplitz structure

The block-Toeplitz Toeplitz-blocks (bTTb) structure arises e.g when modeling 2D homogeneous spatially-invariant phenomena on a rectangular grid. The Kronecker and bTTb structures are related, but not equivalent. An insight is given in the following lemma.

Lemma 7. Let $X \in \mathbb{R}^{N^2 \times N^2}$.

- If X is symmetric block-Toeplitz, then X has a Kronecker rank of N .
- If X has a Kronecker rank of 1, it doesn't *in general* imply neither that X is block-Toeplitz nor has Toeplitz-blocks.

Enforcing the coefficient matrices to be Toeplitz fastens up the identification: at each alternating step, the unknown Toeplitz is embedded into a circulant matrix which is then diagonalized using the Discrete Fast Fourier transform.

7.2 Sparsity

The graphs corresponding to real-networks are sparse, Leskovec et al. (2010). We assume that a node is con-

nected to a limited number of other nodes in the network relatively to the network's size, and therefore we seek to induce sparsity in the matrices $M(a_i^{(j)})$ and $M(b_i^{(j)})$. When the influence of neighboring nodes decay with the distance, a multi-banded structure is equivalent to a banded structure of each factor matrix. When the graph doesn't exhibit such regularity, one other possibility is to induce zero-elements in the estimated parameter vectors $a_i^{(j)}$, $b_i^{(j)}$. Let $\|\cdot\|_0$ denote the zero-norm, then using again λ as regularization parameter, the following constrained optimization problem results,

$$\min_{a_i^{(j)} \in \mathcal{T}_a, b_i^{(j)} \in \mathcal{T}_b} \sum_{k=p+1}^{N_t} \|S(k) - \sum_{i=1}^p \tilde{M}_{a_i} (I_r \otimes S(k-i)) \tilde{M}_{b_i}\|_F^2 + \lambda \|\theta(a_i^{(j)}, b_i^{(j)})\|_0 \quad (25)$$

with

$$\theta(a_i^{(j)}, b_i^{(j)})^T = \left[\dots \left(a_i^{(j)}\right)^T \dots \left(b_i^{(j)}\right)^T \dots \right]$$

In order to preserve the biconvex nature the zero norm is replaced by the 1-norm as follows,

$$\min_{a_i^{(j)} \in \mathcal{T}_a, b_i^{(j)} \in \mathcal{T}_b} \sum_{k=n+1}^{N_t} \|S(k) - \sum_{i=1}^n \tilde{M}_{a_i} (I_r \otimes S(k-i)) \tilde{M}_{b_i}\|_F^2 + \lambda \|\theta(a_i^{(j)}, b_i^{(j)})\|_1 \quad (26)$$

We cannot guarantee theoretically the convergence of this Alternating Sparse Least Squares because the proof of Theorem 6 in Appendix relies on a closed-form expression of each update. Optimization of sparse regularized least squares as in (26) has been widely studied in the literature, see e.g Donoho and Tsai (2006).

7.3 Sequentially Semi-Separable (SSS) matrices

The so-called SSS structure enables standard matrix computations (\times, \cdot^{-1}) to be done in linear computational complexity with respect to the matrix size. For example, inverting a matrix $M \in \mathbb{R}^{N^2 \times N^2}$ written as $M = M_1 \times M_2$ in which both M_1 , M_2 have a SSS structure requires $\mathcal{O}(N)$ operations instead of $\mathcal{O}(N^6)$. Low-rank off-diagonal blocks of the factor matrices that are sought after can be enforced via nuclear norm regularization, see Siquin and Verhaegen (2016) for more details.

8. NUMERICAL EXAMPLES

The proposed *quarks identification* method is now illustrated with a real-life example of a network with unknown communication links. The atmospheric turbulence is a stochastic process that has been modeled via state-space and VAR models or simplified into diagonal VAR models, Correia et al. (2014). When a lightbeam with a flat wavefront passes through a turbulent medium, the wavefront gets distorted. The spatial covariance of the wavefront is not sparse and hence there exists multiple connections from a subsystem to the other in the network. In this paper the turbulence is modeled with a state-space model following the method in Beghi et al. (2008) for one single layer. The quarks identification is performed with the ALS algorithm, and the number of iterations is limited to 10.

The Variance Accounted For (VAF) between two signals y and \hat{y} is defined with:

$$\text{vaf}(y, \hat{y}) = \max\left(0, \left(1 - \frac{\frac{1}{N_t} \sum_{k=1}^{N_t} \|y(k) - \hat{y}(k)\|_2^2}{\frac{1}{N_t} \sum_{k=1}^{N_t} \|y(k)\|_2^2}\right) \times 100\right)$$

Two signals with a VAF equal to 100% are identical. The performance criteria are both the VAF and relative Root-Mean-Square-Error (RMSE) between the signals $S(k+1)$ and $\sum_{i=1}^p \hat{A}_i S(k-i)$.

8.1 Fixed size, varying Kronecker rank and VARX order.

An array of 10×10 phase points is considered. The identification set contains 5.10^3 temporal measurements. In Figure 4 and 5 the coefficient matrix A_1 is displayed when identifying a VARX model with $p = 1$ resp. for the case of estimating the VARX coefficient matrices via least-squares with ℓ_1 sparsity regularization and with KrARX model. In Figure 6 is plotted the relative RMSE in box-plots for different methods these different methods. The patterns of coefficient-matrices identified differ from the method used, and although the sparsity constraint minimizes the number of non-zero entries, it remains detrimental to the overall performance quality with respect to the number of parameters needed to construct the matrix itself. Let us define a measure that we call *model complexity* as the number of non-zero entries needed to construct the p coefficient-matrices. For example, the complexity of a KrARX model is at most $2prN^2$ -only the non-zero elements of the factor matrices-, while it reaches a total of pN^4 for the full least squares estimation. It is illustrated in Figure 7 that displays the VAF with respect to the 0-norm of the entries needed to construct the full coefficient matrix. The VAF obtained with the sparse identification decreases with increasing regularization parameter. The *exact* number of non-zero entry decreases only for high prior on sparsity, for which the VAF is already 0.

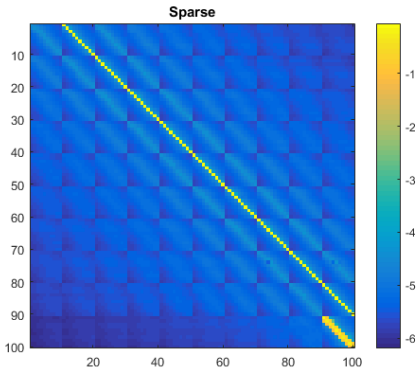


Fig. 4. Coefficient-matrix when identifying a VAR1 model with a regularized sparse least squares. Entries in \log_{10} .

8.2 Scalability

This paragraph investigates how the Kronecker rank of the coefficient-matrices evolves with increasing network size. A total of 7500 points is considered in the identification batch. 10 different experiments were carried out and the Kronecker rank of a VAR1 model ranged from 1 to 4. Figure 8 displays the relative RMSE as a function of the

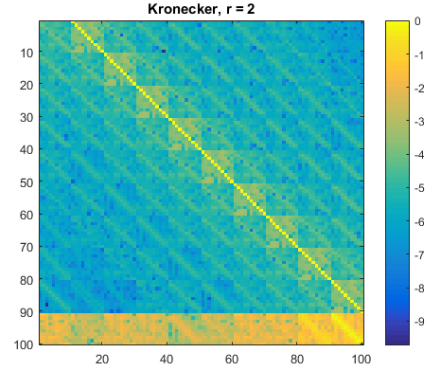


Fig. 5. Coefficient-matrix when identifying a KrARX ($p = 1$) model with a Kronecker structure, $r = 2$. Entries in \log_{10} .

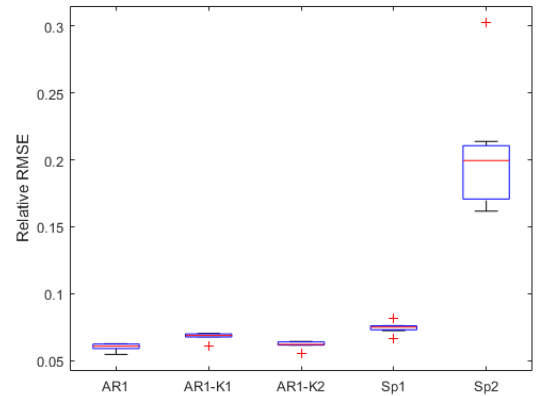


Fig. 6. Relative RMSE for different methods. Sp1 and Sp2 correspond to regularized sparse estimations of the coefficient-matrix A . Sp2 corresponds to sparser model than Sp1.

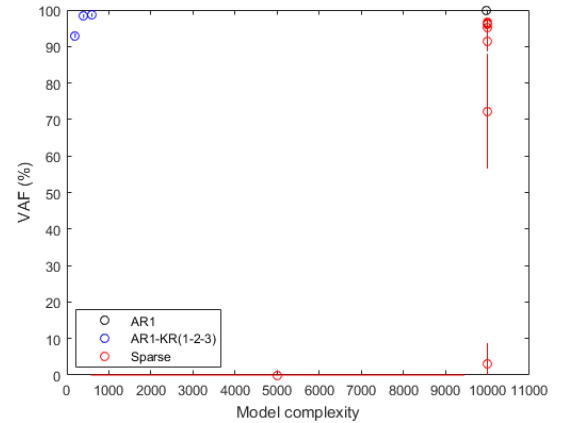


Fig. 7. VAF versus complexity of model. Here, we consider the 0-norm of the coefficient matrix. A blue cross corresponds to a Kronecker rank of resp. 1,2,3. Each red cross corresponds to a regularization parameter on sparsity. The length of the lines are equal to the standard deviation.

total number of phase points, i.e N^2 . The relative RMSE decreases with increasing network size for all methods which is due to the fact that the frozen flow assumption shifts the turbulence phase from one column at every

time sample and the shifted values -easily predictable- constitute the biggest part of the matrix. The ratio of new wavefront entries at each time step to the total number of phase points in the screen is equal to $\frac{1}{N}$, which is the trend observed in Figure 8.

Moreover, we observe that the gap between the least squares solution and the Kronecker estimation does widen with increasing number of points but still keeps a very reasonable relative RMSE. Figure 9 displays the number

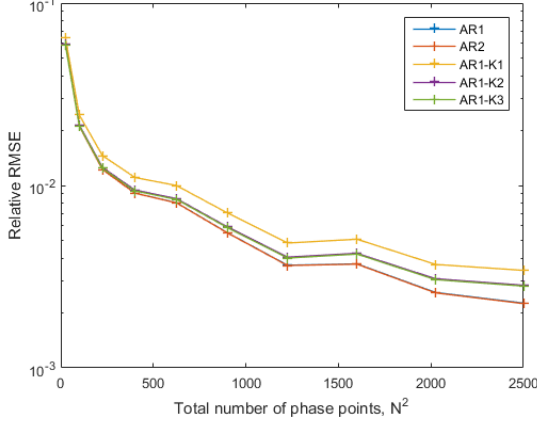


Fig. 8. Relative RMSE for the following methods: AR1-LS, AR2-LS, Kronecker rank 1, 2 and 3. The x-axis is the number of total phase points, and not the number of points in the identification set.

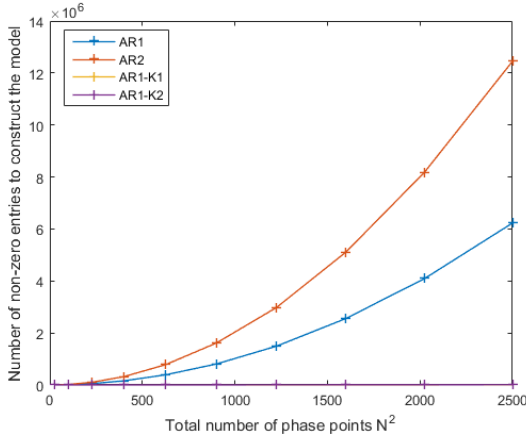


Fig. 9. Model complexity versus size of network

of entries requires to represent the model for both the standard Least-Squares solution and the Kronecker-based representation.

9. CONCLUSION

In this paper is defined the class of Kronecker networks for which the ARX modeling part is investigated. Each coefficient-matrix of the VAR model is approximated with a sum of few Kronecker matrices which offers high data compression for large networks. Estimating in least-squares sense the data matrices give rise to a bilinear problem which is addressed using two methods: first a three-stage non-iterative method is derived, then a iterative alternating least squares whose convergence was proved with a non-zero initial guess. This second framework provides an elegant way of dealing with missing

sensor data and adding further information on the factor matrices, e.g Toeplitz, banded, sparse. Numerical examples on atmospheric turbulence data demonstrates the high compression capabilities of this model as well as its scalability for larger networks. While the ordering of the nodes in the network is crucial for an efficient Kronecker-representation of the coefficient matrix in the VARX model, it may follow the intuition in some cases such as the one presented in the example. The stability of the KVARX model identified has not been presented in this paper and is subject of current investigations.

10. CONVERGENCE PROOF FOR ALS

The proof contains 4 points.

- (1) Positiveness: $\|\mathbf{X}\|_2$ is positive because λ_{max} and the 2-norm for vectors $\|\cdot\|_2$ are both positive.
- (2) If $\|\mathbf{X}\|_2 = 0$, and the matrix $\tilde{\mathbf{U}}$ is full rank, then $\|x\| = 0$ and x is zero. This implies that $\mathbf{X} = 0$ and therefore also the terms $\tilde{\mathbf{U}}(I \otimes x_i)$ for all i .

For the converse, we introduce a partitioning of $\tilde{\mathbf{U}}$ such that $\tilde{\mathbf{U}} = [A_1 \dots A_p]$.

$$\tilde{\mathbf{U}}(I \otimes x_i) = [A_1 x_i \dots A_N x_i] = 0$$

Each block-column A_i is full column rank (persistence of excitation). Hence $x_i = 0$, and $x = 0$.

- (3) Let $\alpha \in \mathbb{R}$.

$$\|\alpha(I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} I \otimes x_1 \\ \vdots \\ I \otimes x_N \end{bmatrix}\|_2^2 = \lambda_{max}(\alpha^2 \tilde{\mathbf{U}}^T \tilde{\mathbf{U}}) \|x\|_2^2$$

Hence,

$$\|\alpha \mathbf{X}\|_2 = |\alpha| \|\mathbf{X}\|_2$$

- (4) Triangular inequality.

$$\begin{aligned} \|\mathbf{X} + \mathbf{Y}\|_2 &= \|(I_N \otimes \tilde{\mathbf{U}}) \left(\begin{bmatrix} I \otimes x_1 \\ \vdots \\ I \otimes x_N \end{bmatrix} + \begin{bmatrix} I \otimes y_1 \\ \vdots \\ I \otimes y_N \end{bmatrix} \right)\|_2 \\ &= \|(I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} I \otimes (x_1 + y_1) \\ \vdots \\ I \otimes (x_N + y_N) \end{bmatrix}\|_2 \\ &= \sqrt{\lambda_{max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})} \|x + y\|_2 \\ &\leq \sqrt{\lambda_{max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})} (\|x\|_2 + \|y\|_2) \\ &\leq \|\mathbf{X}\|_2 + \|\mathbf{Y}\|_2 \end{aligned}$$

$\|\cdot\|_2$ is therefore a norm. \square .

Appendix A. CONVERGENCE PROOF FOR ALS

Proof of Lemma 5

The proof contains 4 points.

- (1) Positiveness: $\|\mathbf{X}\|_2$ is positive because λ_{max} and the 2-norm for vectors $\|\cdot\|_2$ are both positive.
- (2) If $\|\mathbf{X}\|_2 = 0$, and the matrix $\tilde{\mathbf{U}}$ is full rank, then $\|x\| = 0$ and x is zero. This implies that $\mathbf{X} = 0$ and therefore also the terms $\tilde{\mathbf{U}}(I \otimes x_i)$ for all i .

For the converse, we introduce a partitioning of $\tilde{\mathbf{U}}$ such that $\tilde{\mathbf{U}} = [A_1 \dots A_p]$.

$$\tilde{\mathbf{U}}(I \otimes x_i) = [A_1 x_i \dots A_N x_i] = 0$$

Each block-column A_i is full column rank (persistence of excitation). Hence $x_i = 0$, and $x = 0$.

(3) Let $\alpha \in \mathbb{R}$.

$$\|\alpha(I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} I \otimes x_1 \\ \vdots \\ I \otimes x_N \end{bmatrix}\|_2^2 = \lambda_{\max}(\alpha^2 \tilde{\mathbf{U}}^T \tilde{\mathbf{U}}) \|x\|_2^2$$

Hence,

$$\|\alpha \mathbf{X}\|_2 = |\alpha| \|\mathbf{X}\|_2$$

(4) Triangular inequality.

$$\begin{aligned} \|\mathbf{X} + \mathbf{Y}\|_2 &= \|(I_N \otimes \tilde{\mathbf{U}}) \left(\begin{bmatrix} I \otimes x_1 \\ \vdots \\ I \otimes x_N \end{bmatrix} + \begin{bmatrix} I \otimes y_1 \\ \vdots \\ I \otimes y_N \end{bmatrix} \right)\|_2 \\ &= \|(I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} I \otimes (x_1 + y_1) \\ \vdots \\ I \otimes (x_N + y_N) \end{bmatrix}\|_2 \\ &= \sqrt{\lambda_{\max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})} \|x + y\|_2 \\ &\leq \sqrt{\lambda_{\max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})} (\|x\|_2 + \|y\|_2) \\ &\leq \|\mathbf{X}\|_2 + \|\mathbf{Y}\|_2 \end{aligned}$$

$\|\cdot\|_2$ is therefore a norm.

The following lemma corresponds to Lemma 2.2 in Li et al. (2004). It is used to prove that \mathcal{F} is a contraction mapping. It is derived straightforwardly from the definition of the norm above.

Lemma 8. Let $f(\cdot)$ be defined with: $f(\hat{r}) := (M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}$. Under Assumption A2, the magnitude of the directional derivative of $f(\hat{r})$ along a direction vector \mathbf{u} attains its maximum when \mathbf{u} is in the same directions as \hat{r} .

Proof. It is shown that:

$$\|f(\hat{r})\|_2 = \frac{1}{\sqrt{\|M_{\hat{r}}^T M_{\hat{r}}\|_2}}$$

Key in this lemma is therefore to evaluate the term $\|M_{\hat{r}}^T M_{\hat{r}}\|_2$, which we get from the definition of the norm:

$$\|f(\hat{r})\|_2 = \frac{1}{\|\hat{r}\|_2 \sqrt{\lambda_{\max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})}}$$

The proof of the Lemma is therefore unchanged compared with Lemma 2.2 of Li et al. (2004): the contour planes of $f(\hat{r})$ are concentric spheres, and hence the gradient is in the radial direction.

Let us now prove that \mathcal{F} is a contraction mapping on X_L , and that the unique fixed point corresponds to the true parameters.

Proof. The proof consists of first proving that \mathcal{F} is an operator mapping X_L to X_L as $N_t \rightarrow \infty$. Then the operator $\mathcal{F}(\cdot)$ is a contraction mapping on X_L , and last that the unique fixed point is equal to the true parameters r, ℓ .

Note that three norms are used. When x is a vector, $\|x\|_2$ denotes the 2-norm. When $X \in \mathbb{R}^{N_t \times N}$, $\|X\|_2$ is the norm related to the inner product from Definition 4. Else, the induced norm is used. As preliminary, we recall the inequality $\|X\|_2 \leq \|X\|_F$.

We have:

$$\begin{aligned} \hat{L} &= (M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T \tilde{\mathbf{S}} \\ &= (M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T (M_r L + \tilde{\mathbf{V}}) \end{aligned}$$

Therefore:

$$\begin{aligned} \lim_{N_t \rightarrow \infty} \|\hat{L}\|_2 &= \lim_{N_t \rightarrow \infty} \|\mathcal{F}_3(\hat{R})\|_2 \\ &= \lim_{N_t \rightarrow \infty} \|(M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T (M_r L + \tilde{\mathbf{V}})\|_2 \\ &\leq \lim_{N_t \rightarrow \infty} \|(M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T M_r L\|_2 \\ &\quad + \lim_{N_t \rightarrow \infty} \|(M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T \tilde{\mathbf{V}}\|_2 \end{aligned}$$

We start with the right-hand side term:

$$\begin{aligned} \lim_{N_t \rightarrow \infty} \|(M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T \tilde{\mathbf{V}}\|_2 &\leq \lim_{N_t \rightarrow \infty} \frac{\|M_{\hat{r}}\|_2}{\|M_{\hat{r}}^T M_{\hat{r}}\|_2} \|\tilde{\mathbf{V}}\|_2 \\ &\leq \lim_{N_t \rightarrow \infty} \frac{\|\hat{r}\|_2 \sqrt{\lambda_{\max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})}}{\|\hat{r}\|_2^2 \lambda_{\max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})} \|\tilde{\mathbf{V}}\|_2 \\ &\leq \lim_{N_t \rightarrow \infty} \frac{1}{\|\hat{r}\|_2 \sqrt{\lambda_{\max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}})}} \|\tilde{\mathbf{V}}\|_2 \end{aligned}$$

$\tilde{\mathbf{V}}$ has a finite spatial and temporal variance, hence $\lim_{N_t \rightarrow \infty} \|\tilde{\mathbf{V}}\|_2$ is finite. Since $\lim_{N_t \rightarrow \infty} \lambda_{\max}(\tilde{\mathbf{U}}^T \tilde{\mathbf{U}}) = +\infty$,

$$\lim_{N_t \rightarrow \infty} \|(M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T \tilde{\mathbf{V}}\|_2 = 0$$

Therefore,

$$\begin{aligned} \lim_{N_t \rightarrow \infty} \|\hat{L}\|_2 &\leq \lim_{N_t \rightarrow \infty} \|(M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T M_r L\|_2 \\ &\leq \lim_{N_t \rightarrow \infty} \frac{\|M_{\hat{r}}^T M_r\|_2}{\|M_{\hat{r}}^T M_{\hat{r}}\|_2} \|L\|_2 \\ &\leq \lim_{N_t \rightarrow \infty} \frac{\|r^T \hat{r}\|_2}{\|\hat{r}^T \hat{r}\|_2} \|\text{vec}(L)\|_2 \end{aligned}$$

$\frac{\|r^T \hat{r}\|_2}{\|\hat{r}^T \hat{r}\|_2}$ is equal to 1 if and only if $r = \hat{r}$. Else, use can be made of Cauchy-Schwarz inequality and the fact that R is in X_R to prove that it is inferior to 1. We conclude this first part with:

$$\|\hat{\ell}\|_2 \leq \|\ell\|_2$$

Therefore, $\mathcal{F}(\cdot)$ goes from X_L to X_L .

Let us now study the contraction mapping, and determine an upper bound on Q . We have, for $N_t \rightarrow \infty$:

$$\hat{L} = (M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T M_r L$$

Multiplying by $M_{\hat{r}}$ on both left sides gives:

$$\begin{aligned} M_{\hat{r}} \hat{L} &= (M_{\hat{r}}^T M_{\hat{r}})^{-1} M_{\hat{r}}^T M_r L \\ &= M_r L \end{aligned}$$

The right-hand side term reads:

$$M_r L = (I_N \otimes \tilde{\mathbf{U}}) \begin{bmatrix} I \otimes r_1 \\ \vdots \\ I \otimes r_N \end{bmatrix} [\ell_1 \dots \ell_N]$$

and hence, for all i, j :

$$\tilde{\mathbf{U}}(I \otimes r_i) \ell_j = \tilde{\mathbf{U}}(I \otimes \hat{r}_i) \hat{\ell}_j$$

$\tilde{\mathbf{U}}$ being full column rank, we have:

$$\begin{aligned} (I \otimes r_i) \ell_j &= (I \otimes \hat{r}_i) \hat{\ell}_j \\ r_i \ell_j(k) &= \hat{r}_i \hat{\ell}_j(k) \end{aligned}$$

Therefore, since $\ell_j(k) \in \mathbb{R}$,

$$\|r_i\|_2 \|\ell_j(k)\| = \|\hat{r}_i\|_2 \|\hat{\ell}_j(k)\|$$

We have that $\|\hat{r}_i\|_2 = \|r_i\|_2$, which then enables to prove that $\|\ell_j\|_2 = \|\hat{\ell}_j\|_2$. A similar reasoning holds to prove that $\|\hat{r}_{op,i}\|_2 = \|r_i\|_2$.

We now introduce the quantity $Q = \left\| \frac{d\mathcal{F}}{d\hat{L}} \right\|_2$. We have:

$$\left\| \frac{d\mathcal{F}}{d\hat{L}} \right\|_2 \leq \left\| \frac{d\mathcal{F}_v}{d \text{vec}(\hat{L})} \right\|_2$$

where $\mathcal{F}_v(\cdot) = \text{vec}(\mathcal{F}(\cdot))$ is the composition between the vectorized operator and $\mathcal{F}(\cdot)$. From $\hat{L}(k+1) = \mathcal{F}_3(\mathcal{F}_2(\mathcal{F}_1(\hat{L}(k))))$, we decompose:

$$\begin{aligned} Q &= \left\| \frac{d\mathcal{F}}{d\hat{R}} \frac{d\hat{R}}{d\hat{R}_{op}} \frac{d\hat{R}_{op}}{d\hat{L}} \right\|_2 \\ &\leq \left\| \frac{d\mathcal{F}_{3,v}}{d\hat{r}} \right\|_2 \left\| \frac{d\mathcal{F}_{2,v}}{d\hat{r}_{op}} \right\|_2 \left\| \frac{d\mathcal{F}_{1,v}}{d\hat{\ell}} \right\|_2 \end{aligned}$$

Let us start the analysis with $\left\| \frac{d\mathcal{F}_{3,v}}{d\hat{r}} \right\|_2$, and by introducing a deviation from r , denoted with Δr :

$$\left\| \frac{d\mathcal{F}_{3,v}}{d\hat{r}} \right\|_2 = \lim_{\|\Delta R\|_2 \rightarrow 0} \frac{\|\mathcal{F}_{3,v}(\hat{R} + \Delta R) - \mathcal{F}_{3,v}(\hat{R})\|_2}{\|\Delta r\|_2}$$

When $\Delta r \rightarrow 0$, the difference is approximated with:

$$\mathcal{F}_{3,v}(\hat{R} + \Delta R) - \mathcal{F}_{3,v}(\hat{R}) = \text{vec}\left((M_{\hat{r}}^T M_r)^{-1} M_{\Delta r}^T M_r L\right)$$

Therefore,

$$\begin{aligned} \left\| \frac{d\mathcal{F}_{3,v}}{d\hat{r}} \right\|_2 &\leq \left\| \frac{M_{\Delta r}^T M_r}{\Delta r} \right\|_2 \frac{\|\ell\|_2}{\|M_{\hat{r}}^T M_r\|_2} \\ &\leq \frac{\|M_{\Delta r}^T M_r\|_2}{\|M_{\hat{r}}^T M_r\|_2} \|\ell\|_2 \\ &\leq \frac{\|\vec{r}^T r\|_2}{\|\hat{r}^T \hat{r}\|_2} \|\ell\|_2 \end{aligned}$$

where $\vec{r} = \frac{\Delta r}{\|\Delta r\|_2}$. Moreover,

$$\left\| \frac{d\hat{r}}{d\hat{r}_{op}} \right\|_2 = \frac{\|r\|_2}{\|\hat{r}_{op}\|_2} = 1$$

from the update rule $\hat{r}_i(k) = \hat{r}_{op,i}(k) \frac{\|r_i\|_2}{\|\hat{r}_{op,i}(k)\|_2}$. The relationship between the norms of Δr and \hat{r} (which are vectors in the same direction) is $\|\vec{r}^T r\|_2 \|r\|_2 = \|\hat{r}^T r\|_2$. Hence:

$$\begin{aligned} Q &\leq \frac{\|\vec{r}^T r\|_2}{\|\hat{r}^T \hat{r}\|_2} \|\ell\|_2 \frac{\|\vec{\ell}^T \ell\|_2}{\|\hat{\ell}^T \hat{\ell}\|_2} \|r\|_2 \\ &\leq \frac{\|\hat{r}^T r\|_2}{\|\hat{r}^T \hat{r}\|_2} \frac{\|\hat{\ell}^T \ell\|_2}{\|\hat{\ell}^T \hat{\ell}\|_2} \\ &\leq 1 \end{aligned}$$

again using Cauchy-Schwarz and the fact $R \in X_R, L \in X_L$. It proves that \mathcal{F} is a contraction map on X_L .

REFERENCES

Beghi, A., Cenedese, A. and Masiero, A. Stochastic realization approach to the efficient simulation of phase screens. *J. Opt. Soc. Am. A*, volume 25, pages 515–525, 2008.

- Bullmore, E., and Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, volume 10, 2009.
- Correia, C., Jackson, K., Veran, J-P., Andersen, D., Lardi re, O. and Bradley, C. Static and predictive tomographic reconstruction for wide-field multi-object adaptive optics systems. *J. Opt. Soc. Am. A*, volume 31, 2014.
- Donoho, D., and Tsaig, Y. Fast Solution of 1-Norm Minimization Problems when the Solution May Be Sparse. *preprint*, <http://www.stanford.edu/tsaig/research.html>, 2006.
- Granas, A., and Dugundji, J. Fixed Point Theory. *Springer-Verlag, New York*, 2001.
- Kamm, J., and Nagy, J. Optimal Kronecker product approximation of block Toeplitz matrices. *Siam J. Matrix Anal. Appl.*, volume 22, pages 155–172, 2000.
- Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C. and Ghahramani, Z. Kronecker Graphs: An Approach to Modeling Networks. *J. of Machine Learning Research*, volume 11, pages 982–1042, 2010.
- Li, G., Wen, C. and Zhang, A. Fixed point iteration in identifying bilinear models. *Systems & Control Letters*, volume 83, pages 28–37, 2015.
- Loan, C.F. van,. The ubiquitous Kronecker product. *Journal of Computational and Applied Mathematics*, volume 123, pages 85–100, 2000.
- Loan, C.F. van, and Vokt J.P. Approximating matrices with multiple symmetries. *Siam J. Matrix Anal. Appl.*, volume 36, pages 974–993, 2015.
- Loan, C.F. van, and Pitsianis, N.P. Approximation with Kronecker products. *Linear Algebra for Large Scale and Real Time Applications*, Kluwer Publications, Dordrecht, volume 25, pages 293–314, 1992.
- Massioni, P. Distributed control for alpha-heterogenous dynamically coupled systems. *Systems & Control Letters*, volume 72, pages 30–35, 2014.
- Rice, J.K. Efficient algorithms for distributed control: a structured matrix approach. *PhD thesis, Technische Universiteit Delft*, 2010.
- Shen, X., Diamond, S., Gu, Y., Boyd, S. Disciplined Multi-Convex Programming. <http://stanford.edu/boyd/papers/pdf/dmcp.pdf>, 2016.
- Sinquin, B., and Verhaegen, M. Subspace Identification of 1D spatially-varying systems using Sequentially Semi-Separable matrices. *Proceedings of the American Control Conference*, pages 54–59, 2016.
- Tsiligkaridis, T. and Hero, A.O. Covariance Estimation in High Dimensions Via Kronecker Product Expansions. *IEEE Transactions on Signal Processing*, volume 61, 2013.
- Tyrtysnikov, E. Kronecker-product approximations for some function-related matrices. *Linear Algebra and its Applications*, volume 379, pages 423–437, 2004.
- Udell, M., Horn, C., Zadeh, R. and Boyd, S. Generalized Low Rank Models. *Foundations and Trends in Machine Learning*, volume 9, pages 1–118, 2016.
- Wang, J., Zhang, Q., Ljung, L. Revisiting Hammerstein System Identification through the Two-Stage Algorithm for Bilinear Parameter Estimation. *Automatica*, volume 45, no. 11, pages 2627–2633, 2001.
- Zorzi, M., and Chiuso, A. Sparse plus Low rank Network Identification: A Non-parameteric Approach. *Automatica - accepted*, 2016.